Time and Space in Linguistics: Interdisciplinary Computational Approaches & Cross-Creole Comparisons
AARHUS UNIVERSITY, JANUARY 15-16, 2014

Quantitative experiments with the basic vocabulary of the Gulf of Guinea Creoles

Michel Généreux
Tjerk Hagemeijer
(Centro de Linguística da Universidade de Lisboa)

CLUL

Summary

- **Introduction to the Gulf of Guinea Creole (GGC) corpora.**
- **Measuring the orthographic distance between the basic vocabulary of the four GGCs.**
- **Measuring the phonological distance between the basic vocabulary of the GGCs.**
- **Reconstructing the basic vocabulary from the proto-language related to the four creoles using a computer program developed by Oakes (2000) which seeks to implement the Comparative Method.**
- **The Santome and Angolar corpora on the CQPweb.**

CLUL

Project: *The origins and development of creole societies in the Gulf of Guinea: An interdisciplinary study*

http://www.gulfofguineacreoles.com/

The goal of this project is to obtain a composite picture of the origins of the creole populations of the Gulf of Guinea (GG) islands São Tomé, Príncipe and Annobón during the early settlement and consolidation stages (1493-1600) by using an interdisciplinary method with a strong focus on linguistics.

- Lung'Ie (LU)
- Santome (ST)
- Angolar (AN)
- Fa d'Ambô (FA)

- (Computational) Linguistics
- History
- Genetics
- Anthropology

CLUL

The Gulf of Guinea Creole corpora (1)

- The primary goal of the corpora is to carry out comparative research in order to reconstruct lexical and grammatical features of the proto-language.

- All the data in the Gulf of Guinea Creole (GGC) corpora were standardized for spelling in accordance with ALUSTP, a phonology-oriented writing proposal whose main principle is a one-to-one phoneme-grapheme correspondence.

The Gulf of Guinea Creole corpora (2)
Statistics

	tokens		types	sentences	files
	spoken	written			
Angolar	10406	-	992	817	17
Fa d'Ambô	54197	9226	3531	3461	29
Lung’Ie	13149	1305	1321	1206	18
Santome	103655	113311	6787	15188	314
Total	181407	123842	12631	20672	378

CLUL

Ortography-based comparison

We use dynamic programming to compute the edit-distance between two words. This is the minimum number of operations (*deletion*, *substitution*, *insertion*) required to align two words together. This is called the *cost* or the *distance.* Reminder: the same writing system for all the creoles.

For example, in order to transform the Angolar <pupuka> 'to fly' into the unrelated Santome word <vwa> 'to fly', it is necessary to align the two words as follows:

p → 0 (del), u → v (subs), p → 0 (del), u → w (subs), k → 0 (del)

The cost or distance is 5 operations.

```
p u p u k a
| | | | | |
0 v 0 w 0 a
```

CLUL

Excerpt of the extended Swadesh list (orthography)

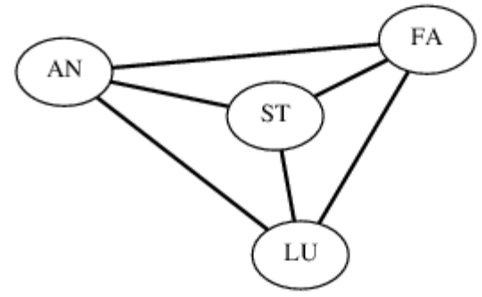
English	Santome	Lung'le	Fa d'Ambô	Angolar
back	<tlaxi>	<taaxi>	<lomba>	<nthusi>
bad	<bluku>	<buuku>	<buuku>	<buuku>
beard	<beba>	<bweba>	<balba>	<fantxi>
because	<punda>	<pidi>	<pokê>	<punda>
belly	<bega>	<bwega>	<beega>	<bega>
bird	<bisu>	<pasu>	<patu>	<situ>
black	<pletu>	<peetu>	<peetu>	<peetu>
blood	<sangi>	<isengi>	<sangi>	<thangi>
bone	<oso>	<osu>	<oso>	<otho>

CLUL

Average distance per word-pair

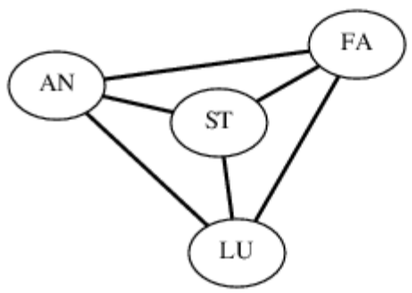
REDUCED is the list without African items (Bantu), which are exclusive to Angolar.

FULL (216)



Word-pair	FULL	REDUCED
AN-FA	2.50	2.10
AN-LU	2.17	1.79
AN-ST	2.00	1.49
FA-LU	1.83	1.75
FA-ST	1.50	1.41
LU-ST	1.17	1.10

REDUCED (163)



CLUL

Phonology-based Comparison

This time we used edit-distance to compute distance based on a phonological model and a larger set of operations. Note that the differences between orthography and phonology are not significant, because of the phonological writing system.

Cost 1: fortition, lenition, aphaeresis, prosthesis, apocope, cluster reduction, syncope, excrescence, epenthesis, vowel break, fusion, assimilation, dissimilation

Cost 2: substitution, deletion, insertion

For ex., it costs 7 to align Angolar /pupuka/ with Santome /vwa/ ‘to fly’.

a → 0 (apocope), uk → wa (assimilation), p →0 (deletion), u → v (substitution), p → 0 (aphaeresis) and only 1 to align the cognates /fi7/ (AN) to /fy7/ ‘cold’ (LU, ST).

```
p   u   p   uk   a          f   i7
|   |   |    |   |          |   |
0   v   0   wa   0          f   y7
```

cost = 1 (assimilation)

CLUL

Excerpt of the phonetic model

Phonetic	Orthographic	Place	Manner	Syllabic	Voice	Nasal	Retroflex	Lateral
/a/	<a>	central, low	-	1	1	0	0	0
/b/	<b>	bilabial	stop	0	1	0	0	0
/d/	<d>	alveolar	stop	0	1	0	0	0
/D/	<dh>	dental	fricative	0	1	0	0	0
/1/	<dj>	post-alveolar	affricate	0	1	0	0	0
/e/	<e>	front, low-mid	-	1	1	0	0	0

Excerpt of the Swadesh list (phonetics)

English	Santome	Lunɑ'Ie	Fa d'Ambô	Anɑolar
back	/tlaxi/	/taaxi/	/lomba/	/nɞusi/
bad	/bluku/	/buuku/	/buuku/	/buuku/
beard	/beba/	/bweba/	/balba/	/fanɲi/
because	/punda/	/pidi/	/pok**2**/	/punda/
belly	/bega/	/bwega/	/beega/	/bega/
bird	/bisu/	/pasu/	/patu/	/situ/
black	/pletu/	/peetu/	/peetu/	/peetu/
blood	/sangi/	/isengi/	/sangi/	/**ɞ**angi/
bone	/oso/	/osu/	/oso/	/oɞo/

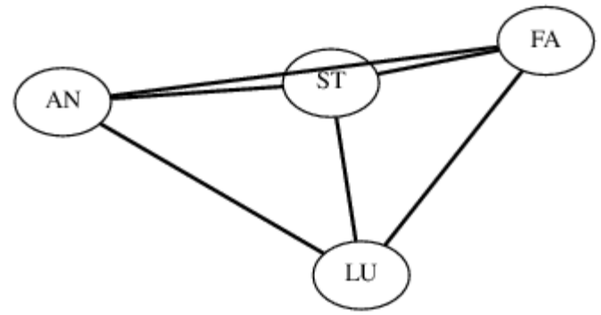
CLUL

Average distance per word-pair

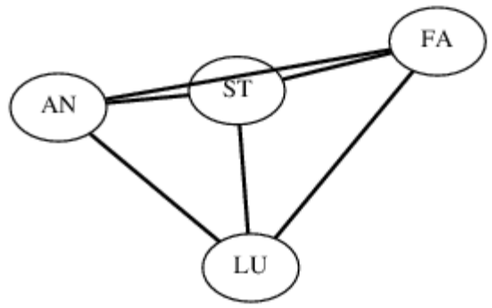
Word-pair	FULL	REDUCED
AN-FA	3.23	2.60
AN-PR	2.66	1.93
AN-ST	2.42	1.54
FA-PR	2.30	2.26
FA-ST	1.91	1.81
PR-ST	1.53	1.40

Angolar is further apart because of the specific Bantu lexicon (only included in the full Swadesh list)

FULL



REDUCED



Reconstructing core vocabulary of the proto-GGC (using the FULL Swadesh list)

We adapted a computer program (Oakes, 2000, “Computer Estimation of Vocabulary in a Protolanguage from Words Lists in Four Daughter Languages”) which implements the “Comparative Method”.

The method follows three main steps:

1) Align word-pairs and record bilingual sound changes in cognates. Collate them statistically;
2) Find regular sound changes in the four creoles, creating a list of proto-phonemes;
3) Substitute phonemes in daughter languages for the corresponding phonemes in the proto-language, if possible.

1) Align word-pairs and record bilingual sound changes. Collate them statistically.

Read the Swadesh 6 word-pair lists (AN-FA, AN-PR, AN-ST, FA-PR, FA-ST and PR-ST) and tally bilingual sound changes for cognates. Word pairs are considered cognate if cost of alignment is less than 5. Here is an excerpt for the pair AN-FA:

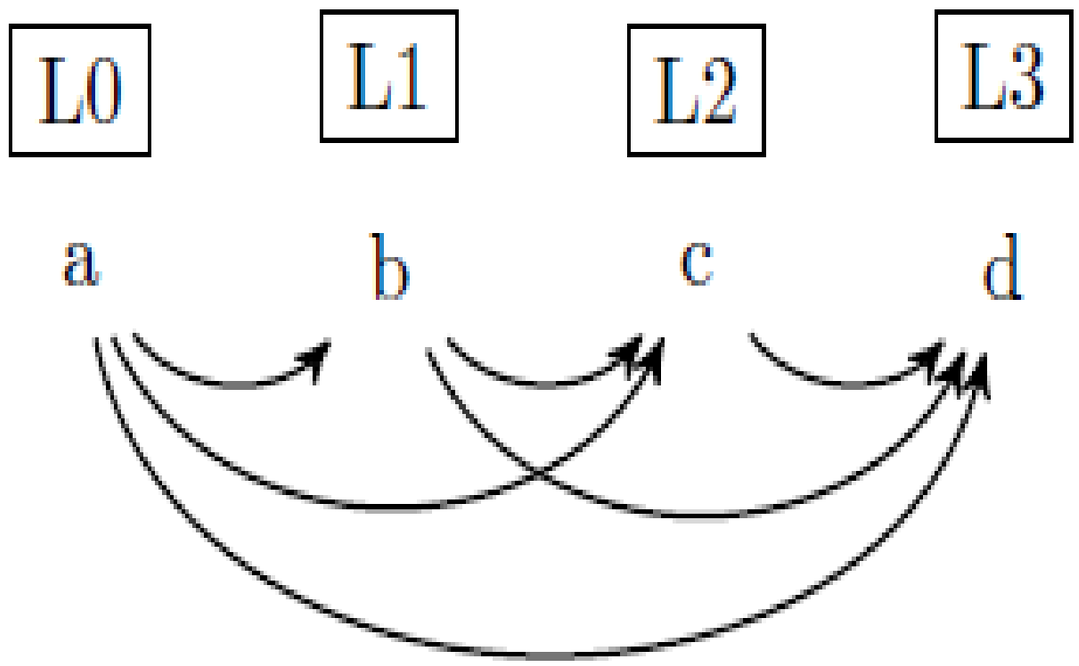
- 7 -> 7 : 10 matches
- 7 -> o : 2 substitutions
- 0 -> n : 6 insertions, apocope, prosthesis
- 9 -> 9 : 3 matches
- D -> z : 2 substitutions
- a -> a : 63 matches

Sound changes occurring less than twice are discarded.

2) Find regular sound changes spanning the four creoles, creating a list of proto-phonemes.

"The technique is to look for instances where L0 and L1 have a regular change a → b, L0 and L2 have a regular change a → c, L0 and L3 have a regular change a → d, L1 and L2 have a regular change b → c, L1 and L3 have a regular change b → d and L2 and L3 have a regular change c → d. If all six conditions hold true, then L0, L1, L2 and L3 exhibit the regular sound change a → b → c → d." (Oakes, 2000). Excerpt from our PROTO:

AN	FA	LU	ST	PROTO
r	d	d	d	d2
e	ee	e	e	e3
t	t	t	t	t
o	7	o	o	o1
8	s	s	s	s
b	b	v	b	b2



3) Substitute phonemes in daughter languages for the corresponding phoneme in the proto-language, if possible (after alignment).

/turu/	/tudu/	/tudu/	/tudu/	PROTO
u	u	u	u	u2
r	d	d	d	d2
u	u	u	u	u2
t	t	t	t	t

PROTO word = /t u2 d2 u2/ from languages 1234

/bi8u/	/limaya/	/bisu/	/bisu/	PROTO
0	~~a~~	0	0	0
0	~~y~~	0	0	0
u	~~a~~	u	u	u2
8	~~m~~	s	s	s
i	~~i~~	i	i	i3
b	~~l~~	b	b	b1

PROTO word = /b1 i3 s u2 0 0/ from languages = 134

CLUL

Substituting language phonemes from corresponding proto-phonemes.

/laba/	/laba/	/lava/	/laba/	PROTO
a	a	a	a	a
b	b	v	b	b2
a	a	a	a	a
l	l	l	l	l

PROTO word = /l a b2 a/ from languages 1234

/b2/	/b2/	/v2/	/b2/	PROTO
2	2	2	2	2
b	b	v	b	b2

PROTO word = /b2 2/ from languages 1234

Some figures

# creoles for Reconstruction	# Reconstructions
2	75
3	56
4	39
Total	170
No Reconstruction	46

Creoles	# Reconstructions involved
AN	122
FA	96
LU	125
ST	147

Evaluation for Proto-Malyo-Javanic creoles	
Identical to Nothofer	At least one morpheme is different
13	42

CLUL

Fine-tuning (work in progress)

- Accuracy of proto-phonemes needs further assessment

- The proposed proto-words always have a reflex in one or more creoles but are generally accurate (conservative outcome)

- Some forms are wrongly excluded as cognates and require the inclusion of language-specific regular sound changes in order to improve accuracy

- Results are expected to improve further if a larger amount of cognates can be made available

Conclusion and Future Work

- **AN-FA appear to be the most distant from a phonological point of view (orthographically too)**

- **LU-ST appear to be the least distant from a phonological point of view (orthographically too)**

- **AN, FA and LU are closer to ST than to each other, confirming the central role of ST within the language group**

- **Reconstruction of 79% of the core vocabulary of the four GGCs, but fine-tuning and evaluation are required**
  - **Examine the list of frequent sound changes to see if new regular patterns apply to the GGCs**
  - **Use the distance measure to extract additional potential cognates in Gulf of Guinea data**

CLUL

Santome and Angolar on the CQPweb (1)

A sample from the Santome is available on a web platform for searching and making concordances:

- 17261 tokens in 137 texts
- manually annotated with POS
- metadata: type, language, genre, author, source, date, title, notes, age, place
- http://alfclul.clul.ul.pt/CQPnet/coforro/

A sample from the corpus Angolar is available on a web platform for searching and making concordances:

- 10253 tokens in 15 texts
- manually annotated with POS
- http://alfclul.clul.ul.pt/CQPnet/angolar/

CLUL

Santome and Angolar on the CQPweb (2)

Your query "[word="tudu" %c]" returned 108 matches in 36 different texts (in 17,261 words [137 texts]; frequency: 6256.88 instances per million words) [0.055 seconds]

|< << >> >| Show Page: 1 Line View Show in random order New query Go!

No	Filename	Solution 1 to 50 Page 1 / 3		
1	sun_tataluga	alê se . Jingantxi se me so tava ka toma konta di	tudu	ben d' alê se . xiga #ua belu dja , Sun Tataluga
2	sun_tataluga	pixkadô , mina di pôvô , fô omali . se ,	tudu	pixkadô ku ba omali na pega nê #ua kaxta di pixi fa
3	sun_tataluga	a lanta Sun Tataluga ba liba , zo komesa ka glita ni	tudu	lwa di poson se : Sun Tataluga - Homem Popular . se
4	sun_tataluga	Sun Tataluga - Homem Popular . se , a fe fesa ni	tudu	xitu . Tataluga so pasa govena tela se non . na pasa
5	pinto_da_costa_sa_ka_tlaba	fa Pinto da Costa sa ka tlaba da Costa sa ka tlaba	tudu	pôvô sa ka bê pôvô nganha kuxensa Santome zud' e da non
6	bon_dja_di_desu	an ê pô lembla di kondê potxi Bondja di Dêsu ê ,	tudu	pôvô bili za , taluvê segu ka bê wê tela dê xinku
7	a_desu_plama_bili_za	klonvesadu ! ! ngê ku ka da non djêlu di kopla	tudu	kwa ku non mêsê , biza ... sa migu di non ,
8	a_desu_plama_bili_za	sun Govenadô manjo Carlo de Souza Gorgulho . bili xtlivisu pa	tudu	ngê tlaba sosegadu sê tlomentu , sê vôlô , dentxi betu so
9	a_desu_plama_bili_za	punda ngê ku na mêsê tlaba fa sa ladlon pligisôzu ... punda	tudu	inen mosu ku fe supetu kyê wê ba xtlivisu . xka balha
10	a_desu_plama_bili_za	sun Govenadô bi Santome , ladlon fia kôkôkô , punda solo ,	tudu	ngê tê xtlivisu dja djingu donu ku tê kwa sa ke .
11	tudu_tela_ku_ua_lungwa_se_me_tan	ku a ka saya olha , kabêsa ka jinga , manda .	Tudu	tela ku #ua lungwa se me tan tudu tela tava ka fla
12	tudu_tela_ku_ua_lungwa_se_me_tan	jinga , manda . Tudu tela ku #ua lungwa se me tan	tudu	tela tava ka fla #ua lungwa se me , tudu inen tava

CLUL